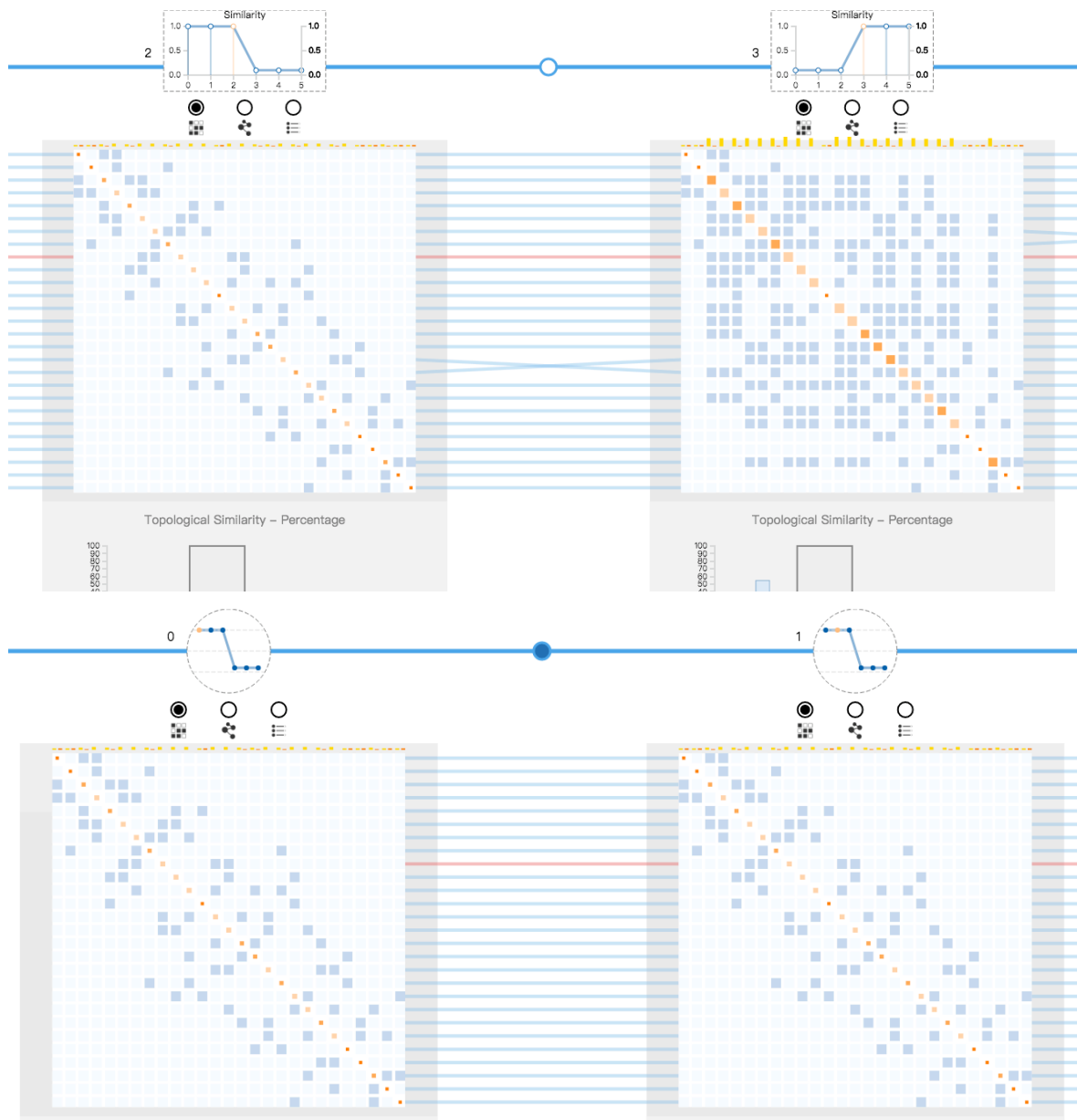


# 2017/11/20-2017/11/26 周报

## DONE

- **考试**：考试于本周一正式结束，还有两门课的课程大作业需要完成。
- **投稿**：根据曹楠老师和郭博的想法，对系统进行了部分修改，分成两个版本，主要区别在于图表的详细情况，两个版本都保留了，供最后选择。



- **论文阅读**：
  - 一篇图数据挖掘综述，[Graph mining: A survey of graph mining techniques](#)，发在ICDIM上，好像在ABC类刊物上都没看到，而且篇幅较短，肯定无法对整个图数据挖掘进行概述（只交代了一些任务：frequent pattern mining, classification, clustering），以及一些重要的论文，比较简单，只作为随意阅览的读物。

- 一篇图数据挖掘异常检测综述, [Graph based anomaly detection and description: a survey](#), 发在DMKD上, 对图数据异常检测进行了一个比较详细的概述, 因为要写课程论文的原因, 只读了其中关于静态图拓扑结构异常检测部分, 以下是阅读的时候做的一些简要的笔记, 在介绍这些方法的时候, 作者讲某一类, 会分成两步: main idea, approaches, 在approaches里面会对一些重要的论文进行一个简述。

对于静态图结构异常的检测 (不包括节点属性), 主要可以分成两大类: 基于结构的手段和基于社团的手段。

- 基于结构的方法

基于结构的方法又可以分为两个大类, 一种基于结构特征, 比如节点度或者子图的中心度; 另一种则是通过图结构来量化节点之间的邻近关系 (通常是最短路径), 从而辨识节点之间的关联。

- 基于结构特征的方法: 通过抽取图的结构特征, 从而可以构造出一个特征空间, 图中的异常检测问题就可以被转化为离群点检测问题。图的结构特征一般有节点特征, 比如出度入度, 中心度, 邻近度等等; 二元特征 (两个节点之间的), 比如边中间度, 公共邻接点数等等; 中心网络特征, 比如三角数量, 总权重, 主要特征值等; 节点组特征, 比如紧密度, 密度, 模块度等等; 以及整图特征, 比如连通子图数, 子图大小分布等。
- 基于邻近关系的方法: 相邻的节点被认为有相似的分类 (比如, 感染/健康), 一个广泛使用的测量邻近度的方法 (Personalized PageRank) 可以用来测量节点和某个初始点的邻近度 (closeness)。在定义了邻近度之后, 从而就可以判断节点的重要性, 以此来甄别其是否是异常。

- 基于社团的方法

基于社团的方法, 或者说基于聚类的方法, 主要是为了发现高密度的连通社团, 从而发现连接不同社团的桥接边或桥接节点, 一般这些桥接两个紧密社团的边或节点会被认为是异常。基于社团的方法主要需要解决两个问题, 一是如何找到一个给定点所在的社团, 也就是如何找到节点的邻域; 而是如何确定一个给定的节点是否是一个桥接节点。

对于第一个问题, 一般会用PPR分数 (Personalized PageRank) 来度量所有其他节点和给定节点之间的邻近度, 从而构造出节点的邻域; 对于第二个问题, 每个节点根据它的邻域内节点的PPR分数的均值, 计算该节点的正常度。

- 一篇2010年的关于静态图拓扑结构异常检测的论文 (为了撰写数据挖掘的课程论文读的, [OddBall: Spotting Anomalies in Weighted Graphs](#)), 文章的检测对象是以某个节点为中心的egonet, 先定义了异常模式 (一般的egonet会符合某些规则, 而异常的则不符合), 作者罗列了四条规则, 用以度量中心节点和周围节点的关系, 主要用egonet的边数量、邻接节点数量、整体权重和邻接矩阵的特征值来衡量。对于每条规则, 作者计算了每个egonet的分数, 将异常检测问题转化成outlier检测的问题。
- 一篇挖掘相似子图的图数据挖掘论文 ([Fast Similar Subgraph Search with Maximum Common Connected Subgraph Constraints](#)), 是清华大学的一篇论文, 发在BigData Congress上, 主要是基于最大公共连接子图, 来快速搜索在整张图里面与给定的子图, 相似的子图。相似度计算, 是利用最大公共连接子图在两个图中的占比来进行衡量, 占比接近的两个子图, 相似度更高。

**TODO**

- 软件注册书撰写
- 课程作业：数据挖掘论文、三维CAD建模的project

任务	截止日期	当前进度
RCAalyzer文章投稿	10月底	详见上方
大图可视化调研	9、10月份	正在寻找合适的技术框架、系统
关于palantir软件注册撰写	11月底	未开始
硕士论文	待确定	已经完成开题报告，还差论文主体部分